Dynamic and Static congestion models: A review

André de Palma, Mogens Fosgerau

ECOLE POLYTECHNIQUE

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

DYNAMIC AND STATIC CONGESTION MODELS: A REVIEW

André DE PALMA
Mogens FOSGERAU

November, 2010

Cahier n° 2010-28

DEPARTEMENT D'ECONOMIE
Route de Saclay
91128 PALAISEAU CEDEX
(33) 1 69333033
http://www.economie.polytechnique.edu/
mailto:chantal.poujouly@polytechnique.edu

Dynamic and static congestion models: a review[i]

André de Palma
Ecole Normale Supérieure de Cachan and Ecole Polytechnique

Mogens Fosgerau
Technical University of Denmark

November, 2010

We begin by providing an overview of the conventional static equilibrium approach. In such model both the flow of trips and congestion delay are assumed to be constant. A drawback of the static model is that the time interval during which travel occurs is not specified so that the model cannot describe changes in the duration of congestion that result from changes in demand or capacity. This limitation is overcome in the Vickrey/Arnott, de Palma Lindsey bottleneck model, which combines congestion in the form of queuing behind a bottleneck with users' trip-timing preferences and departure time decisions. We derive the user equilibrium and social optimum for the basic bottleneck model, and explain how the optimum can be decentralized using a time-varying toll. They then review some extensions of the basic model that encompass elastic demand, user heterogeneity, stochastic demand and capacity and small networks. We conclude by identifying some unresolved modelling issues that apply not only to the bottleneck model but to trip-timing preferences and congestion dynamics in general

Introduction

This paper provides a brief introduction to dynamic congestion models, based on the Vickrey (1969) bottleneck model which has become the main workhorse model for economic analysis of situations involving congestion dynamics.

The word *dynamic* can have several possible meanings. One possibility is that it relates to the way traffic systems evolve and users learn from day to day. In the context of the bottleneck model, it relates to intra-day timing, i.e. to the interdependencies between traffic congestion at different times within a given day.

We shall discuss dynamic approaches against the background of static models. Static models assume that congestion is constant over some given time period. A congestion law provides the travel time as a function of the entering flow. The time dimension is not explicitly involved: all quantities are computed as single figures specific to a time period.

The basic static model considers a network comprising nodes and links. The nodes are centroids of zones, associating trip ends within a zone with a point that is a node in the network. Links connect the nodes. A cost function describes the cost of using each link. Congestion means that the cost increases as the number of users of the link increases. The demand is given by the origin-destination (O-D) matrix, indicating the number of trips between pairs of nodes. The solution involves the choice of route within the network for each O-D pair. Traffic volume on each link, the travel cost of using each link, the cost of making each trip, and the total travel cost for all users all depend on these route-choice decisions.

Each user for each O-D pair is assumed to choose a route in the network that minimises the sum of link costs for the trip. But users compete for the same space and the route choices of users in one O-D pair affect the costs experienced by other users through congestion. We can imagine a process where users keep revising their route choices in response to the route choices of other users. We seek an equilibrium in which no user can reduce his cost by choosing a different route. This equilibrium concept is due to Wardrop (1952). This problem was first given a mathematical formulation and solution for a general network by Beckmann, McGuire and Winston (1956). We will discuss the static model in more detail below in the context of simple networks.

The static model remains a basic tool for the mathematical description of congested networks. The static model does, however, omit important features of congestion. The static model is hence unsatisfactory for a number of purposes.

The main feature that the static model omits is that congestion varies over the day, with pronounced AM and PM peaks in most cities. Travel times can easily increase by a factor two from the beginning to the height of the peak. To design and evaluate policies for tackling congestion it is necessary to recognize these variations. There are a number of fundamental features of congested demand peaks that a model should take into account.

First, travellers choose not only a route, but also a departure time in response to how congestion varies over the course of a day. When a policy is implemented that affects peak congestion, travellers respond by changing departure time. The departure time changes are systematic on average and can be observed in the aggregate temporal shape of the peak. Think, for example, of car traffic entering the central business district (CBD) of some large city. The number of travellers reaching their workplaces per hour is fixed at the capacity rate during the morning peak. So if the

number of workplaces in the CBD increases, the duration of the morning peak must increase too. Similarly, if capacity is increased the duration of the peak will shrink. The duration of the peak thus depends on both demand and capacity. Such observations suggest that trip timing is endogenous and speak in favour of dynamic models.

Second, travellers incur more than just monetary costs and travel time costs when they make a trip. Travellers have preferences regarding the timing of trips and deviations from the preferred timing are costly. Such scheduling costs are comparable in magnitude to congestion-delay costs as a fraction of total user costs. These scheduling costs are by nature ignored in static models. This means that static models cannot reveal the effect of policies that affect scheduling costs.

Third, many relevant policies can only be described within a dynamic model. A congestion toll or parking fee that varies over time as congestion increases and decreases is an obvious example.

The basic dynamic model discussed in this paper, the bottleneck model, starts directly from the above observations regarding within-day dynamics. It is therefore well suited to analyse policies that rely on these dynamics. It was introduced by William Vickrey (1969). Arnott, de Palma, and Lindsey (1993a) revisited and extended this seminal but almost forgotten model. It is a tractable model and it leads to a number of important insights. The model features one origin-destination pair (let us say residence and workplace), one route, and one bottleneck. The bottleneck represents any road segment that constitutes a binding capacity constraint. The bottleneck allows users to pass only at some fixed rate. There is a continuum of users and it takes some positive interval of time for them all to pass the bottleneck. Users are identical and they wish to arrive at the destination at the same ideal time t^*. Because of the bottleneck, all but one user must arrive either before or after t^*. Deviation from t^* represents a cost for users. They also incur a travel time cost, which includes free flow travel time and delays in the bottleneck. Individuals choose a departure time to minimise the sum of schedule delay and travel time costs.

To analyse this situation, we consider an equilibrium in which no traveller has incentive to change his departure time choice. This is an instance of a Nash equilibrium (Haurie & Marcotte, 1985), which is the natural generalisation of Wardrop equilibrium. Individuals are identical and therefore they experience the same cost in equilibrium. One might wonder whether the Nash equilibrium concept has any counterpart in the real world. We see Nash equilibrium as a benchmark. Like anything else in our models, it is an idealisation, describing a situation that we hope is not too far from reality. The appeal of Nash equilibrium is that it is a rest point for any dynamic mechanism whereby informed travellers revise their (departure time) choice, if they do not achieve the maximum utility available to them.

Travellers incur the same generalised travel cost in equilibrium, but they have different trips. Some depart early, experience only a short delay at the bottleneck, but arrive early at work. Others avoid queuing delay by departing late, but arrive also late at work. Those who arrive near the preferred arrival time will experience most congestion and have the longest travel time. In this way, the bottleneck model describes a congested demand peak with a queue that first builds up and then dissolves.

The endogenous choice of the departure time was independently studied by de Palma, Ben-Akiva, Lefèvre & Litinas (1983), who proposed a dynamic model incorporating a random utility departure time choice model and a generalised queuing model. In contrast to the Vickrey bottleneck

model, where the capacity constraint is either active or not, the supply model of de Palma et al. shifts smoothly from the uncongested to the congested regime.

An area of economic literature has grown out of these two initial contributions, exploring a number of issues in the context of the basic bottleneck model: e.g., equilibrium, social optimum, decentralization of the social optimum via pricing, second best pricing (including step tolls), elastic demand, heterogeneous individuals, small networks (routes in parallel and routes in series), stochastic capacity and demand, alternative treatments of congestion, and pricing on large networks. The basic model has also been extended to include mode choice, parking congestion, modelling of the evening commute and non-commuting trips. The research stream initiated with M. Ben-Akiva had more focus on numerical computation, and it has led, amongst other development, to the METROPOLIS software for large networks, discussed below.

This paper first reviews the simple static model of congestion, where time is not explicitly considered. This serves as a background for the dynamic model. We then introduce the basic bottleneck model and continue to discuss some of the extensions mentioned.

The static model of congestion[ii]

Static networks

We begin with a simple example. Consider a fixed number $N > 0$ of travellers having two routes available. The travellers split with $n_1 > 0$ on the first route and $n_2 > 0$ on the second route, where $n_1 + n_2 = N$. The cost associated with each route is taken to be a linear function of traffic such that the average cost on route i is $C_i(n_i) = a_i + b_i n_i$. The cost is a so-called generalised cost, combining monetary cost and travel time in a single monetary equivalent. The Nash equilibrium occurs when no traveller wants to change route, which requires that $C_1(n_1) = C_2(n_2)$. Solving this equation leads to the equilibrium solution[iii]

$$n_1^e = \frac{a_2 - a_1}{b_1 + b_2} + \frac{b_2}{b_1 + b_2} N, n_2^e = N - n_1^e .$$

The Nash equilibrium has every traveller minimise his/her own cost. We can alternatively consider social optimum where the total cost for all travellers is minimised. In general the social optimum is not a Nash equilibrium. The social optimum minimises the total cost function

$$\min_{n_1, n_2} W(n_1, n_2) = n_1 C_1(n_1) + n_2 C_2(n_2) .$$

The total cost associated with use of route i is $n_i C_i(n_i)$. The marginal cost of an additional user is

$$\frac{d\left[n_i C_i(n_i)\right]}{dn_i} = C_i(n_i) + b_i n_i .$$

In this expression, $C_i(n_i)$ is the cost paid by the marginal user. The remainder $b_i n_i$ is an externality: it is the part of the increase in the total cost that is not borne by the additional user. The first-order condition for social optimum requires equal marginal costs, or

$$C_1(n_1) + b_1 n_1 = C_2(n_2) + b_2 n_2 . \quad (1.1)$$

The only difference between this and the first-order condition for the equilibrium is the terms representing the externalities of the two routes. The externalities are zero if $b_i = 0$, i=1,2, i.e. if adding an additional user does not lead to increased travel cost. In this case, the social optimum would be the same as the equilibrium.

The social optimum has

$$n_1^o = \frac{a_2 - a_1}{2b_1 + 2b_2} + \frac{2b_2}{2b_1 + 2b_2} N,\ n_2^0 = N - n_1^o . \qquad (1.2)$$

The solution is written in this way to emphasise the similarity to the Nash equilibrium. The only difference between the optimum and the equilibrium outcomes is that the marginal costs, the b_i, have been replaced by $2b_i$ in the expression for the optimum outcome. This indicates that the optimum can be achieved as an equilibrium outcome by setting a toll equal to $n_i b_i$ on each of the two routes. This has the effect of doubling the variable cost from the perspective of users and the expression in (1.2) then becomes the equilibrium outcome.

Elastic demand

The discussion so far has considered a fixed number of travellers N. We now allow demand to be elastic, limiting attention to just one route. Travellers on this route are identical, except for different willingness to pay to travel. Figure 1 shows a downward-sloping inverse demand curve $D(N)$ to reflect that demand decreases as the cost increases. The curve $C(N)$ is again an average cost curve expressing the cost that each traveller incurs. The curve $MC(N)$ is a marginal cost curve, expressing the marginal change in total cost following a marginal increase in the number of travellers; in other words[iv]

$$MC(N) = C(N) + N \cdot C'(N).$$

When the cost curve is increasing, the marginal cost curve will lie above the cost curve.

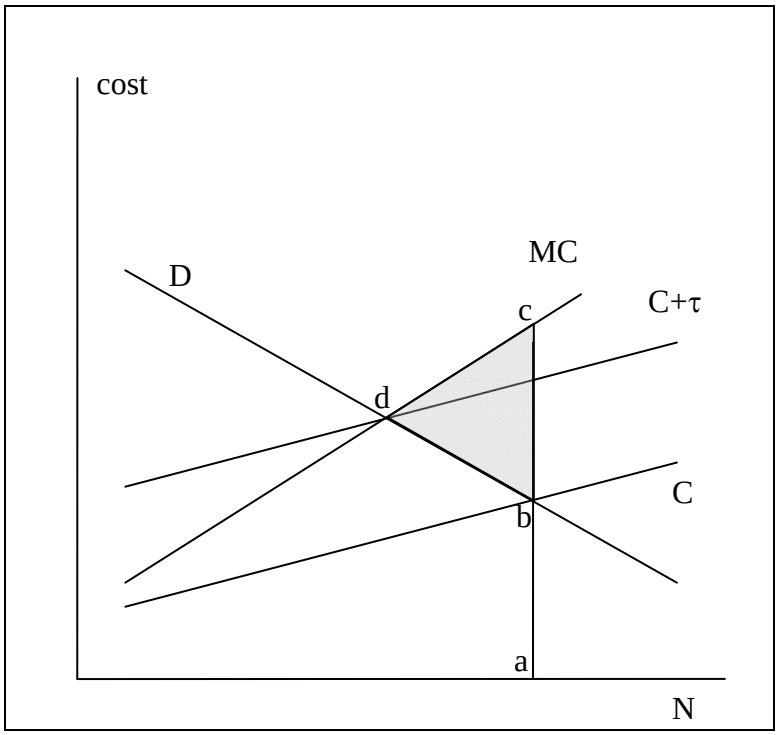

Figure 1. A static model

The equilibrium occurs at the intersection of the demand curve with the average cost curve at the point b. The marginal traveller at this point is indifferent between travelling and not travelling, he faces a cost corresponding to the line segment a-b and a benefit of the same size. For travellers in aggregate, however, the cost of adding the marginal traveller is given by the MC curve. For the marginal traveller at point b, this cost corresponds to the line segment a-c. So the last traveller imposes a net loss corresponding to the line segment b-c on the group of all travellers. If usage was reduced to the point where the MC curve crosses the demand curve, then the corresponding loss is zero for the traveller at the point d. The total loss in market equilibrium is then represented by the shaded triangle b-c-d on the figure.

The optimal toll, labelled τ in Figure 1, implements the optimum at the point d, where the private benefit is equal to the marginal cost. The toll is required because drivers ignore the costs they impose on other drivers. The toll is just the difference, evaluated at the social optimum, between the marginal cost and the average cost, i.e. the externality.

The basic bottleneck model

We now introduce the basic Vickrey bottleneck model in its simplest form. Consider a continuum of $N > 0$ identical travellers, who all make a trip. They have to pass a bottleneck, which is located d_1 time units from the trip origin and d_2 time units from the destination. Denote the time of arrival at the bottleneck of a traveller by t and the exit time from the bottleneck as a. The situation is illustrated in Figure 2. A traveller departs from the origin at time $t - d_1$ and arrives at the

bottleneck at time t. There he/she is delayed until time $a \geq t$ at which time he/she exits from the bottleneck to arrive at the destination at time $a + d_2$.

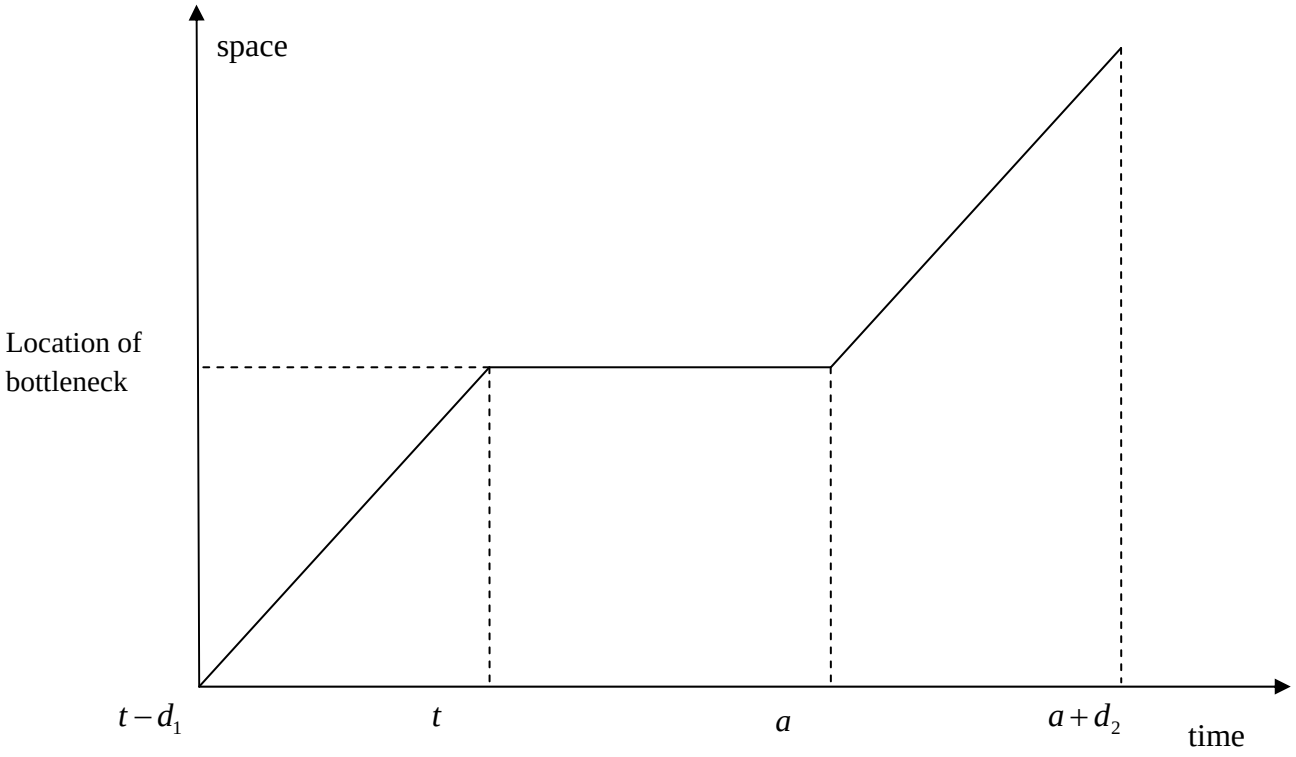

Figure 2. Trip timing

Each traveller has a scheduling cost expressing his/her preferences concerning the timing of the trip. Travellers are assumed to have a preferred arrival time t^* and they dislike arriving earlier or later at the destination. Travellers also prefer the trip to be as quick as possible. For a trip that starts at time t_1 and ends at time t_2, consider then a cost of the form

$$c(t_1, t_2) = \alpha \cdot (t_2 - t_1) + \beta \cdot \max(t^* - t_2, 0) + \gamma \cdot \max(t_2 - t^*, 0), \quad (1.3)$$

where $0 < \beta, 0 < \gamma$ and $\beta < \alpha$. In this formulation, α is the marginal cost of travel time, β is the marginal cost of arriving earlier than the preferred arrival time, γ is the marginal cost of arriving later, and these values are constant. The deviation $t_2 - t^*$ between the actual arrival time and the preferred arrival time is called schedule delay and it is possible to speak of schedule delay early and schedule delay late, depending on the sign of the schedule delay.[v] This cost formulation has become colloquially known as $\alpha - \beta - \gamma$ preferences. Later, we shall consider scheduling cost of a general form.

The travel time d_1 between the origin and the bottleneck adds the same constant amount to the scheduling cost of all travellers and so it can be set to zero without affecting the behaviour of travellers in the model. Similarly, the travel time d_2 between the bottleneck and the destination can

be set to zero by redefining the preferred arrival time. So without loss of generality we may let $d_1 = d_2 = 0$. This means that the time of departure is the same as the time of arrival at the bottleneck and that the time of exit from the bottleneck is the same as the time of arrival at the destination.

Travellers depart from the origin according to an aggregate schedule, described in terms of the cumulative departure rate R, where $R(a)$ is the number of travellers who have departed before time a. So R is similar to a cumulative distribution function: it is proportional to the probability that a random traveller has departed before time a. R is increasing, since travellers never return. Moreover, $R(-\infty) = 0$ and $R(\infty) = N$. The departure rate $\rho(a) = R'(a)$, wherever R is differentiable.

The bottleneck can serve at most s travellers per time unit. Travellers who have not yet been served wait before the bottleneck. The bottleneck serves travellers in the sequence in which they arrived (first-in-first-out or FIFO). The bottleneck capacity is always used if there are travellers waiting before it.

Recall that Nash equilibrium is defined as a situation in which no traveller is able to decrease his cost by choosing a different departure time. Since travellers are identical, this definition reduces to the requirement that all travellers experience the same cost and that the cost would be higher for departure times that are not chosen by any travellers.

Denote the interval of departures and arrivals as $I = [a_0, a_1]$. Let us consider some properties of Nash equilibrium. First, there will be queue from the time the first traveller departs until the last traveller departs, since otherwise there would be a gap in the queue and somebody could move into the gap to decrease cost. Second, the queue will end at the time the last traveller departs, since otherwise he/she could wait until the queue was gone and reduce cost. This shows that the departure interval is just long enough for all travellers to pass the bottleneck. Third, as the cost of the first and the last travellers are equal and since they experience no queue, they must experience the same cost due to schedule delay. These insights are summarised in the following equations.

$$a_1 - a_0 = N / s, \tag{1.4}$$

$$\beta \cdot (t^* - a_0) = \gamma \cdot (a_1 - t^*). \tag{1.5}$$

Equation (1.4) ensures that arrivals take place during an interval that is just long enough that all travellers can pass the bottleneck. Equation (1.5) ensures that no traveller will want to depart at any time outside I.

Solving these two equations leads to

$$a_0 = t^* - \frac{\gamma}{\beta + \gamma} \frac{N}{s},$$

$$a_1 = t^* + \frac{\beta}{\beta + \gamma} \frac{N}{s}$$

and the equilibrium cost for every traveller is

$$\frac{\beta\gamma}{\beta + \gamma} \frac{N}{s} \equiv \delta \frac{N}{s}.$$

This is linear in the number of travellers and so the simple static model could be viewed as a reduced form of the dynamic model.

Equations (1.4) and (1.5) are extremely useful in that they determine the equilibrium cost of travellers as a function of the number of travellers and the bottleneck capacity. The total cost is then $\delta N^2 / s$ with corresponding marginal cost $2\delta N / s$, of which half is internal cost to each traveller and the other half is external. The marginal change in total cost following a change in capacity s is $-\delta N^2 / s^2$. Since there is no toll, price equal travel cost: $p^e = \delta N / s,$ that is price is a function of N and s. The function is this a reduced-form supply function, which is very usefull, especially in analytical work, together with a trip demand function.

There is always a queue during the interval I. This means that the bottleneck capacity is fully utilised and hence that sd travellers pass the bottleneck during an interval of length d. At time a, a total of $R(a)$ travellers have entered the bottleneck, taking a total time of $R(a)/s$ to pass. The first traveller enters and exits the bottleneck at time a_0. Hence a traveller arriving at bottleneck at time a exits at time $a_0 + R(a)/s$. Travellers are identical so they incur the same scheduling cost in equilibrium. Normalising $t^* = 0$, it emerges that

$$\delta \frac{N}{s} = \alpha \cdot \frac{R(a)}{s} + \beta \max\left(-a_0 - \frac{R(a)}{s}, 0\right) + \gamma \max\left(a_0 + \frac{R(a)}{s}, 0\right).$$

Differentiating this expression leads to

$$\rho(a) = \begin{cases} s\dfrac{\alpha}{\alpha - \beta}, a_0 + \dfrac{R(a)}{s} \le 0 \\ s\dfrac{\alpha}{\alpha + \gamma}, a_0 + \dfrac{R(a)}{s} > 0 \end{cases}$$

during interval I. A few observations are immediately available. Initially the departure rate is constant and higher than s (since $\beta < \alpha$). It is high until the traveller who arrives exactly on time. Later travellers depart at a constant rate which is lower than s.

Figure 3 shows the resulting departure schedule. The horizontal axis is time and the vertical axis is the number of departures, ranging from 0 to N. The thick kinked curve is the cumulative departure rate R. Departures begin at time a_0 and end at time a_1 with $R(a_1) = N$. The line segment connecting point a_0 to point e represents the number of travellers served by the bottleneck, it has slope s.

The first departures take place at a rate larger than capacity and queue builds up. For example, at time a, the number of travellers who have departed corresponds to the length of the segment $a - c$, while the number of travellers who have been served by the bottleneck corresponds to the length of the segment $a - b$. Thus the queue at that time has length corresponding to the segment $b - c$. The travellers in the queue at time a will all have been served by time d, which is then the time at which the traveller departing at time a is served by the bottleneck. The time spent in the bottleneck equals the length of the queue at the time of departure divided by the capacity.

The traveller departing at time d exits the bottleneck exactly at time a_*. Therefore the departure rate drops below capacity at this time and the queue begins to dissolve. It also follows that the queue reaches its maximum length at time d.

For the top half of the figure, the horizontal time axis refers both to the departure from the origin and to the arrival time at the destination. For the bottom half of the figure, the time axis instead refers to the arrival time at the destination. The shaded areas on the bottom half of Figure 3 show the composition of the scheduling cost throughout the peak. The first traveller arrives early and is not delayed in the bottleneck so his cost is $\beta \cdot (a_* - a_0)$. Later travellers do not arrive as early, but are delayed more in the queue and incur the same trip cost. The traveller who arrives at the preferred arrival time is the most delayed and his trip cost comprises solely travel time cost. Later arrivals are less delayed in the queue, but arrive later at the destination. The last traveller is not delayed in the bottleneck at all, but arrives last at the destination and incurs a cost of $\gamma \cdot (a_1 - a_*)$.

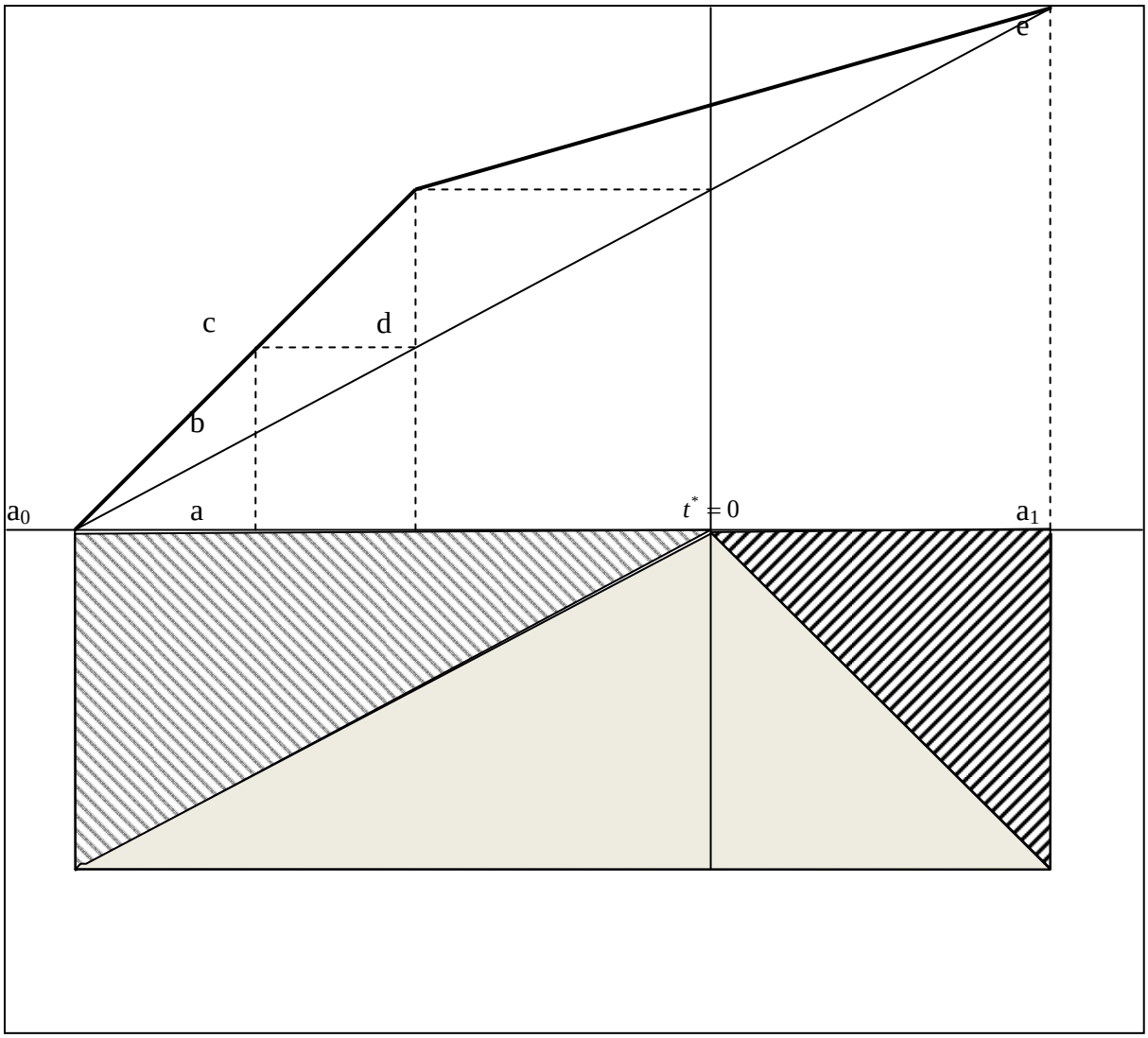

Figure 3. Equilibrium departure schedule under $\alpha - \beta - \gamma$ preferences

Optimal tolling

The queue that arises in equilibrium in the bottleneck model is sheer waste. It generates no benefit at all. If travellers could be induced to depart at the capacity rate s during the equilibrium interval I, then there would be no queue. All travellers (except the very first and the very last) would gain from reduced travel time while arriving at the destination at exactly the same time as in equilibrium. A main insight of the bottleneck model is that it is possible to achieve this outcome through the application of a toll.

So consider a time varying toll $\tau(\cdot) \geq 0$ charged at the time of arrival at the bottleneck. We make the additional behavioural assumption that travellers choose departure time to minimise the sum of the toll and the trip cost. We restrict attention to tolls that have $\tau(a_0) = \tau(a_1) = 0$ and are zero outside the departure interval I. This means that (1.4) and (1.5) still apply. If the toll is well-behaved, in ways to be explained below, then Nash equilibrium exists and departures still occur in the interval I. Therefore the equilibrium cost is the same as in the no-toll equilibrium discussed above.

Travellers do not lose, but somebody else may gain since revenue from the toll can be used for other purposes. The size of the toll revenue is

$$\int_{a_0}^{a_1} \tau(s) ds, \tag{1.6}$$

and this represents a net welfare gain.

Since the cost must be constant in equilibrium, we have

$$\tau(a) = \delta \frac{N}{s} - c\left(a, \frac{R(a)}{s} + a_0\right), \tag{1.7}$$

where R is now the departure rate that results when the toll is imposed. It is (intuitively) clear that maximal efficiency is attained when the toll revenue is as large as it can be without destroying the equilibrium. Increasing $\tau(a)$ in (1.7) will reduce $R(a)$.[vi] Moreover, the queue cannot be negative and so we must require that $R(a) \geq s(a - a_0)$. Therefore the maximal toll maintains zero queue and the least possible cumulative departure rate, i.e. $R(a) = s \cdot (a - a_0)$. This corresponds to a constant departure rate $\rho(a) = s$. The optimal toll is

$$\tau(a) = \delta \frac{N}{s} - c(a, a) = \delta \frac{N}{s} - \beta \cdot \max(-a, 0) - \gamma \cdot \max(a, 0)$$

for $a \in I$ and zero otherwise. This toll is initially zero at time a_0. Then it increases at the rate β until it reaches a maximum of $\delta N / s$ at time 0. It then decreases at the rate γ until it is again zero at time a_1. The optimal toll corresponds to the grey shaded area in Figure 3. In a sense, it just replaces the cost of queueing by a toll. The efficiency gain is achieved because queueing is pure waste whereas the toll revenue is just a transfer.

Elastic demand

The discussion of the bottleneck model so far has assumed demand to be inelastic. A natural extension is to assume that the number of travellers deciding to participate in the peak depends on the equilibrium cost (Arnott et al., 1993a). The trip cost

$$p = \tau(a) + c\left(a, a_0 + R(a)/s\right) \tag{1.8}$$

is the same for all travellers in equilibrium. This implies that the total toll payment is $N \cdot (p - \bar{c})$, where $\bar{c}$ is the average scheduling cost of travellers. Let $N(\cdot) > 0, N'(\cdot) < 0$ be a downward sloping demand function such that $N(p)$ is the realised demand.

This is a very convenient way to extend the model: conditional on any equilibrium number of travellers, the properties of equilibrium are exactly the same as in the inelastic case. The equilibrium number of travellers is uniquely determined since demand is decreasing as a function of the equilibrium cost of travellers while the equilibrium cost of travellers is increasing as a function of the number of travellers. This simplicity comes, however, at a cost as it requires separability between trip timing on the one hand and participation on the other.

The separability of trip timing and participation implies that the optimal toll with elastic demand is the same as in the case of inelastic demand. To see this, note first that the optimal toll is able to remove queuing, so the average cost of travellers remains equal to $\delta N / s$. Consider the following welfare function

$$W(p) = \int_p^\infty N(s)\,ds + N \cdot (p - \bar{c}),$$

i.e. the sum of consumer surplus and the total toll revenue. To find the welfare optimising toll, note that

$$\bar{c} = \frac{s}{N} \int_{a_0}^{a_1} c(a, a)\,da,$$

which can be shown to imply that

$$\frac{\partial \bar{c}}{\partial p} = \frac{N'(p)}{N(p)} \left(\delta N / s - \bar{c}\right).$$

Using this to evaluate the first-order condition for maximum of $W(p)$ leads to $p = \delta N / s$. That is, the optimal price should equal the equilibrium scheduling cost. Using (1.8) shows that the optimal toll is $\tau(a) = \delta N / s - c(a, a)$, which is the same as in the case of inelastic demand.

Optimal capacity and self-financing

Consider now a situation in which the optimal toll applies while capacity s is supplied at cost $K(s) \geq 0$, with $K' > 0$. We extend the social welfare function with the cost of capacity provision

$$W(p, s) = \int_p^\infty N(r)\,dr + N \cdot (p - \bar{c}) - K(s).$$

For any given capacity s, the optimal value of $\tau(a) = \delta N / s - c(a, a)$ is as shown above. Note that

$$\frac{\partial \bar{c}}{\partial s} = \frac{1}{s} \left(\delta N / s - \bar{c}\right).$$

This can be used to show that capacity is optimal when $sK'(s) = N \cdot (p - \overline{c})$. That is, the revenue from the optimal toll is equal to $sK'(s)$.

This finding leads directly to the self-financing theorem for the bottleneck model. If capacity is produced at constant returns to scale, i.e. if $K(s) = sK'(s)$ with $K'(s)$ constant, then the optimal toll exactly finances the optimal capacity $K(s) = N \cdot (p - \overline{c})$. If there are increasing returns to scale, then $K(s) > sK'(s)$, in which case the optimal toll cannot finance the optimal capacity.

The self-financing result is also called the cost recovery theorem. It is an instance of a general self-financing theorem by Mohring & Harwitz (1962), which assumes that travel cost is homogenous of degree zero in capacity and use. A number of results on self-financing are summarised by Verhoef & Mohring (2009).

The optimal capacity can be computed in the three regimes: no toll, coarse step toll and optimal fine toll. It can be shown that the optimal capacity is the lowest for the optimal fine toll, intermediary for the coarse toll and larger for the no toll regime (see ADL, 1993, for a proof). Note that these proofs are correct with inelastic (and elastic) demand.

Scheduling preferences

General formulation

The $\alpha - \beta - \gamma$ formulation of scheduling cost used above is a special case of more general scheduling preferences, introduced in this section. Below we revisit the bottleneck model from the perspective of these general scheduling preferences.

In order to describe the traveller choice of trip timing in a more general way, we formulate scheduling preferences for a given trip in the form of scheduling utility $u(t_1, t_2)$, where t_1 is the departure time and t_2 is the arrival time,. We shall make minimal assumptions regarding the specification of u.

It is natural to require that $u_1 = du / dt_1 > 0$, such that it is always preferred to depart later, given t_2.[vii] Similarly, requiring $u_2 = du / dt_2 < 0$ ensures that arriving earlier is always preferred, given t_1. A marginal increase in travel time then always leads to a utility loss, since travellers will either have to depart earlier or arrive later. Define the function $v(a) = u(a, a)$ as the scheduling utility that a traveller would receive if travel was instantaneous. Assume that v is quasi-concave and attains maximum at $v(t^*)$. This assures that for any $d > 0$ there is a unique solution to the equation $v(a) = v(a + d)$. It also implies that v is increasing for $a < t^*$ and decreasing for $a > t^*$.

We incorporate monetary cost by considering utility to be $u - \tau$. In some cases it is more convenient to talk about cost, which will then be the negative of utility, i.e. $\tau - u$. In either case, it is implied that there is separability between scheduling and monetary cost. That is, a constant cost does not affect the preferences regarding trip timing.

In some situations it is necessary to specify scheduling utility further by imposing a certain functional form. For example, the $\alpha-\beta-\gamma$ formulation specifies the scheduling cost completely up to a few parameters. Such restriction can be necessary for reasons of identification in econometric work, but in general it is preferable to specify as little as possible, since restricting the model entails the risk of introducing errors. In theoretical models it is similarly preferable to work with general formulations, since otherwise there is a risk that the results one may obtain depend on the specific formulation.

In some cases it may be considered acceptable to impose a separability condition, just as we have done in the case of monetary cost and trip timing. The timing of the trip is given by a departure time and an arrival time and we work under the assumption that these times are all that matter about trip timing. The travel time is the difference between the departure time and the arrival time. We could equivalently describe trip timing in terms of travel time and arrival time or in terms of travel time and departure time. From the perspective of general scheduling utility $u(t_1,t_2)$, this leads to three possibilities for introducing a separability condition.

$$u(t_1,t_2)=f(t_2-t_2)+g(t_1)$$
$$u(t_1,t_2)=f(t_2-t_1)+g(t_2)$$
$$u(t_1,t_2)=f(t_1)+g(t_2)$$

The first condition would say that scheduling utility is separable in travel time and departure time. The second condition would say instead that scheduling utility is separable in travel time and arrival time. The $\alpha-\beta-\gamma$ scheduling cost is a special case of this second possibility: Changing the travel time does not affect the traveller preferences regarding arrival time and vice versa. The third possible separability condition is used in the Vickrey (1973) formulation of scheduling preferences that we will consider in the next section. Here scheduling utility is separable in departure time and arrival time. That is, changing departure time, does not affect the preferences regarding arrival time and vice versa.

The concept of the preferred arrival time t^* was used to define the $\alpha-\beta-\gamma$ scheduling cost. It makes sense to talk about a preferred arrival time when there is separability in travel time and arrival time, since then the preferred arrival time is not affected by the travel time. Without this separability, there is no single preferred arrival time since the preferred time to arrive depends on the travel time. If instead scheduling utility is separable in departure time and travel time, then we would want to talk about a preferred departure time. In some contexts, for example the PM commute from work to home, this might be a more natural concept. In general, neither the concept of a preferred arrival time nor a preferred departure time may be relevant. We shall now discuss Vickrey (1973) scheduling preferences, which are separable in departure time and arrival time.

Vickrey (1973) scheduling preferences

Consider an individual travelling between two locations indexed by $i=1,2$. He derives utility at the time dependent rate η_i at location i. Let us say he starts the day at time T_1 at location 1 and

ends the day at time T_2 at location 2. If he departs from location 1 at time t_1 and arrives (later) at location 2 at time t_2, then he obtains scheduling utility

$$u(t_1,t_2)=\int_{T_1}^{t_1}\eta_1(s)\,ds+\int_{t_2}^{T_2}\eta_2(s)\,ds. \tag{1.9}$$

The formulation is illustrated in Figure 4.

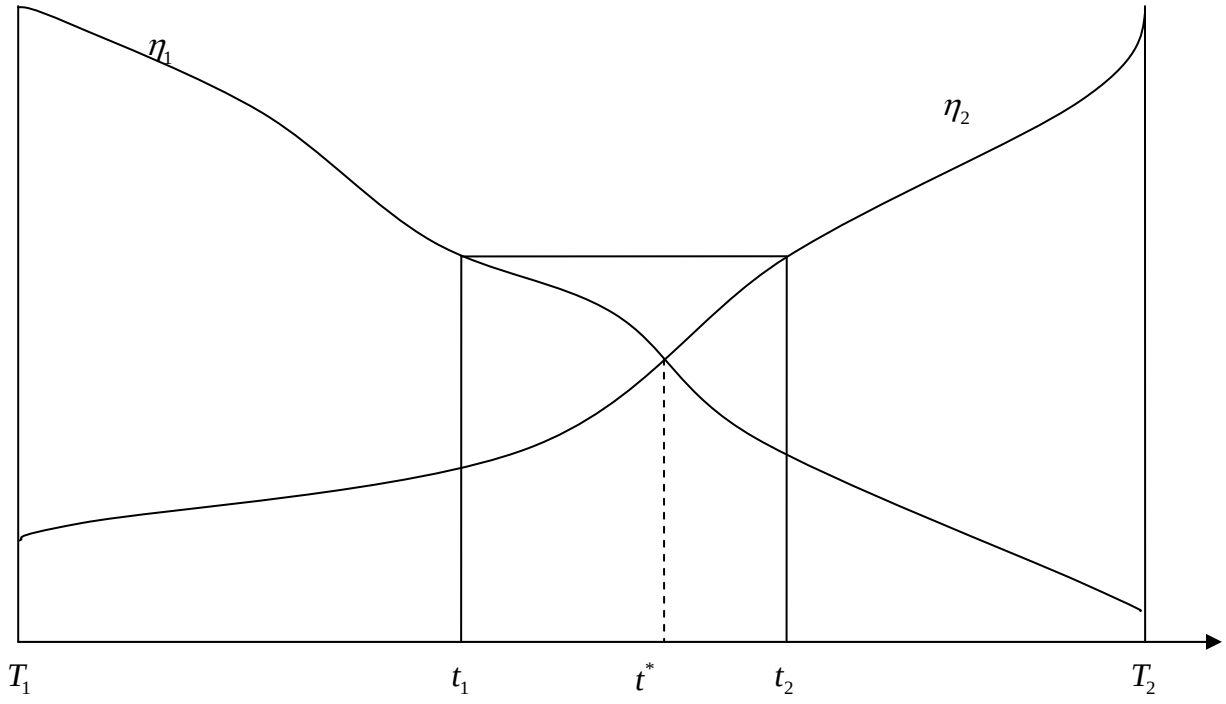

Figure 4. Vickrey (1973) scheduling preferences

Note that when T_1 and T_2 are fixed, these numbers can be replaced by arbitrary numbers in (1.9) without affecting the implied preferences. Assume that $\eta_1>0,\ \eta'_1<0$, $\eta_2>0,\ \eta'_2<0$ and that there is a point in time, t^*, where $\eta_1(t^*)=\eta_2(t^*)$. Speaking in terms of the morning commute these conditions imply that a traveller prefers to be at home or at work to travelling, that his/her marginal utility of staying later at home is decreasing, that his/her marginal utility of arriving earlier at work is also decreasing, and that there is a time (t^*) when he/she would optimally transfer from home to work if instant travel was possible. Given a travel time of d, he/she would optimally depart at the time $t(d)$ depending on d when $\eta_1(t(d))=\eta_2(t(d)+d)$. It is straightforward to derive that his/her value of time would be

$$-\frac{\partial u(t(d),t(d)+d)}{\partial d}=\eta_2(t(d)+d).$$

This is strictly increasing as a function of d. Using survey data on stated choice, Tseng & Verhoef (2008) provide empirical estimates of time varying utility rates corresponding to the Vickrey (1973) model.

The cost of travel time variability

When travel time is random and travellers are risk averse, the random travel time variability leads to additional cost, the cost of travel time variability. Both Vickrey formulations of scheduling preferences are useful for deriving measures of the cost of travel time variability as well as of the scheduling impact of the headway of scheduled services. Such cost measures can be useful to incorporate elements of dynamic congestion in reduced form in static models. Consider a traveller who is about to undertake a given trip. The travel time for the trip is random from the perspective of the traveller. While he/she does not know the travel time outcome before making the trip, the traveller knows the travel time distribution. The travel time distribution is independent of the departure time of the traveller. The latter is a strong assumption but necessary for the results

The traveller is assumed to choose his departure time optimally, so as to maximise his/her expected scheduling utility. That makes the expected scheduling utility a function just of the travel time distribution. Therefore it is possible in principle to evaluate how the expected scheduling utility depends on the travel time distribution. Simple expressions are available for the two Vickrey specifications of scheduling preferences.

In the case of $\alpha-\beta-\gamma$ preferences, Fosgerau & Karlstrom (2010) show that the expected trip cost with optimal departure time is

$$\alpha\cdot\mu+\sigma\cdot(\beta+\gamma)\int_{\gamma/(\beta+\gamma)}^{1}\Phi^{-1}(s)\,ds,$$

which is linear in the mean and in the standard deviation of travel time. This is a practical advantage in applications. The expression depends on the shape of the travel time distribution through the presence of Φ in the integral and so Φ must be taken into account if the marginal value of standard deviation of travel time is to be transferred from one setting to another. In the same vein, Fosgerau (2009) uses $\alpha-\beta-\gamma$ scheduling cost to derive simple expressions for the value of headway for scheduled services. In the case of Vickrey (1973) scheduling preferences with linear utility rates, Fosgerau & Engelson (2010) carry out a parallel exercise. They show that with random travel time and unconstrained choice of departure time, the expected scheduling cost with the optimal choice of departure time is linear in travel time, travel time squared and the variance of travel time. Parallel results are also provided for the value of headway for scheduled services. In contrast to the case of $\alpha-\beta-\gamma$ scheduling cost, it is possible also to derive a simple expression for the expected scheduling cost for the case of a scheduled service with random travel time.

The bottleneck model revisited

The results discussed above for the basic bottleneck model survive in some form with more general scheduling preferences. The setup of the model is as before, the only change is that now travellers are only assumed to have scheduling preferences of the general form discussed above.

Without loss of generality we may again consider $d_1 = d_2 = 0$, since the exact form of scheduling preferences is not specified.

It is easy to argue, using the same argument as in the simple case, that Nash equilibrium requires departures in an interval $I = [a_0, a_1]$ satisfying

$$a_1 - a_0 = N / s, \tag{1.10}$$

$$v(a_0) = v(a_1). \tag{1.11}$$

This is illustrated in Figure 5. Moreover, the queue has length zero at time a_0 and a_1 but it is strictly positive at any time in the interior of this interval. The second condition (1.11) has a unique solution since v is quasiconcave and it ensures that no traveller will want to depart at any time outside I.

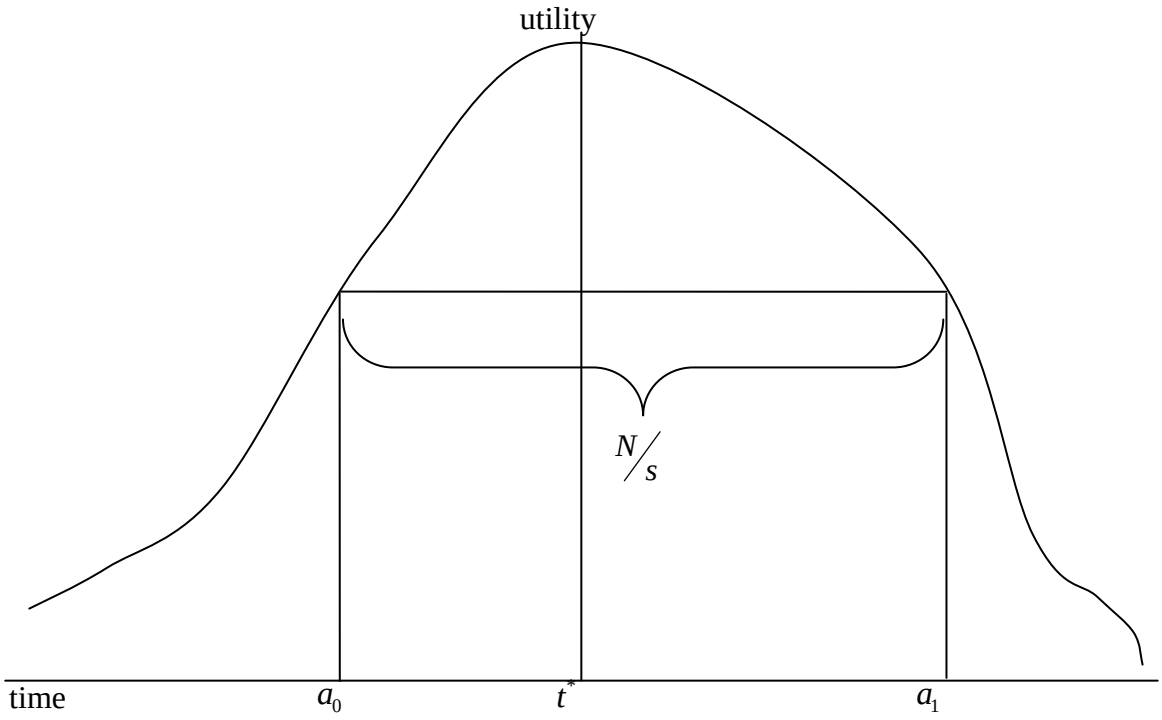

Figure 5. The function v and the equilibrium departure interval

Equations (1.10) and (1.11) determine the equilibrium utility of travellers as a function of the number of travellers and the bottleneck capacity. It is then straightforward to derive the marginal external congestion cost and the marginal benefit of capacity expansion.

As in the basic model, there is always a queue during the interval I and a traveller arriving at the bottleneck at time a exits at time $a_0 + R(a)/s$. Travellers are identical so they achieve the same scheduling utility in equilibrium

$$v(a_0) = u\left(a, a_0 + \frac{R(a)}{s}\right).$$

Consider now a time varying toll $\tau(\cdot) \geq 0$ charged at the time of arrival at the bottleneck. We restrict attention to tolls that have $\tau(a_0) = \tau(a_1) = 0$ and are zero outside the departure interval I. This means that equations (1.10) and (1.11) still apply. If the toll is not too large, then Nash equilibrium exists with departures still in the interval I.[viii] Therefore the equilibrium utility $v(a_0)$ is the same as in the no-toll equilibrium. As in the basic model, the optimal toll maintains the departure rate at capacity. The optimal toll is then given by $\tau(a) = v(a) - v(a_0)$ for $a \in I$ and zero otherwise.

The conclusions regarding elastic demand extend to the case of general scheduling preferences. That is, the optimal toll is still $p = \tau - u$, which is the same as in the case of inelastic demand. The conclusions regarding optimal capacity and self-financing also carry over to the general case. That is, if capacity is supplied at constant cost and optimally chosen, then the optimal toll exactly finances the capacity cost.

Extensions of the bottleneck model

The bottleneck model is useful in many ways. It generates a number of insights concerning dynamic congestion, while being still relatively simple and tractable. The model is useful if the mechanisms it describes are representative of the real world. It is, however, a highly stylised description of actual congested networks. It is therefore of interest to extend the model by introducing more relevant features. Such an exercise has two main purposes. One is to gauge the robustness of the conclusions of the basic model. We can have greater confidence in conclusions that survive in more general versions of the model. The other main purpose is to generate new insights that were not available with the basic model. This section proceeds with a presentation of some of the extensions of the bottleneck model available in the literature.

Second best pricing

The optimal toll described above varies continuously over time. A real toll could do the same to any relevant degree of precision, but there remains the problem that travellers may not be able to understand such a complex pricing structure. Moreover, there may be technological reasons for varying tolls less frequently. Acceptability of road pricing is also a fundamental issue (see, e.g. de Palma, Lindsey & Proost, 2007, on this issue).

Such considerations have led researchers to consider tolls that vary in steps. In the context of the bottleneck, ADL (1990) consider the simplest step toll, namely a toll that is positive and constant during some interval and zero otherwise. Such a toll has also been called a coarse toll.

The discrete jumps of such a toll generate some new properties of the resulting equilibrium. Three groups of travellers can be identified according to whether they travel before, during or after the tolling period. Figure 6 compares the cumulative departure curve in the step-toll equilibrium and compares it with the no-toll equilibrium. Consider first the time before the toll is turned on. The cost of the last traveller not to pay the toll should be the same as the cost of the first traveller to pay the toll. To achieve this equality, there must be a period with no departures between these two

travellers. Early in the morning travellers depart at a high rate, they pay no toll and consequently depart at the same rate as they would in no-toll equilibrium. Just before the departure time at which travellers would begin to pay the toll, departures cease for a while and the queue dissipates gradually as travellers are served by the bottleneck.

Departures start again when the queue has diminished just enough for the toll payment to be compensated by lower queueing time. The optimal single step toll is timed such that the queue has just disappeared at the time the toll kicks in. Departures for the group of travellers paying the toll then continue following the pattern analysed above. The toll is constant for these travellers and hence does not affect the departure rates. The departure rate is consequently high until the time at which a traveller arrives at the destination exactly on time, and then it drops to a lower level. The optimal single step toll is timed such that the queue has just disappeared at the time the toll lifts.

A new phenomenon emerges relating to the third and final group of travellers who do not pay the toll. As shown in Figure 6, there is no queue at the moment before they depart. But the first traveller to depart must have the same cost in equilibrium as the other travellers. This can only happen if there is a mass departure at this time. In a mass departure, travellers depart so closely together that their sequence in the queue is random. In this case, travellers are assumed to account for their expected trip cost.

It turns out that all the remaining travellers depart at once under the optimal coarse toll if $\alpha < \gamma$ as has been found in most empirical studies. On average they are better off than a traveller who waits until the queue has gone before departing. But all travellers must achieve the same expected cost in equilibrium. Therefore the first traveller departs later under the coarse toll than under no toll.

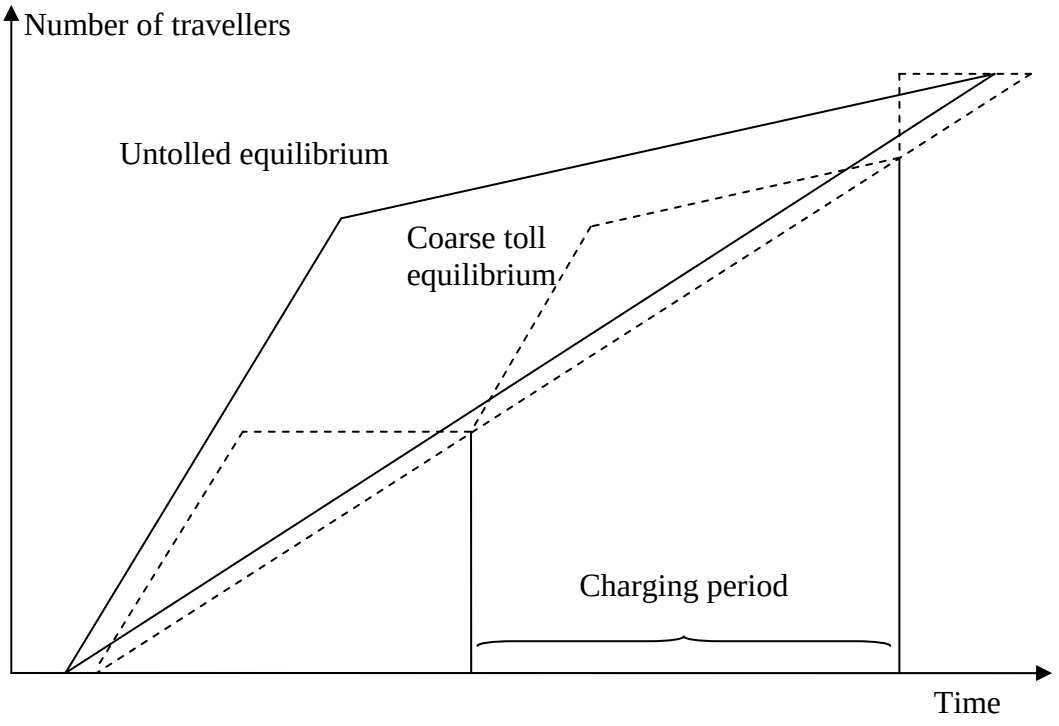

Figure 6. The optimal coarse toll

ADL (1990) carried out their analysis for the case of a single step coarse toll. Laih (1994; 2004) extended this analysis to the case of multistep tolls using a slightly modified queueing technology in which some travellers can wait in a separate queue for the toll to lift, while those paying the toll pass the bottleneck.[ix] Laih then showed that at most n/(n + 1) of the total queueing time can be eliminated with the optimal n-step toll. Daganzo & Garcia (2000) also consider a step toll with the modified queueing technology. They divide travellers into two groups. Travellers from the first group are not liable to pay any toll. Travellers from the second group are liable to pay a constant step toll if they pass the bottleneck during the tolling period, otherwise they do not have to pay any toll. If the toll is high enough then travellers from the tolled group will avoid the tolling period. The tolling period is timed such that it fits exactly with the equilibrium departure interval of the untolled group. As a consequence, travellers from the untolled group can find an equilibrium during the tolling period and be strictly better off than without the scheme. Travellers from the tolled group are not worse off, since they travel during the same interval as without the scheme and avoid paying any toll by travelling outside the toll period. The essential insight is that the equilibrium cost is determined by the first and last travellers (as in (1.5) or (1.11)) as long as capacity is fully utilised during the departure interval.

The function of the toll in this example is to reserve the bottleneck capacity for a specific group of travellers during a specific interval of time. Shen and Zhang (2010) describe a mechanism that uses ramp metering to achieve a similar effect.

Random capacity and demand

ADL (1999) consider bottleneck congestion in a situation where capacity varies randomly from day to day. The ratio is fixed within a day and given the ratio the evolution of the queue is then deterministic. Travellers choose departure time without knowing the random ratio of the day. They are assumed to find equilibrium in expected utility given the information they have. ADL identify circumstances in which the static model is not consistent with a reduced form of the dynamic model. A perhaps surprising result is that providing more information can decrease welfare when demand is elastic and congestion is not efficiently tolled.

Lindsey (2009) considers self-financing in the bottleneck model with random capacity and demand. He finds that the Mohring-Harwitz self-financing theorem survives randomness as long as the information used to set the optimal toll is the same as the information that is available to travellers.

de Palma and Fosgerau (2009) include random travel time variability in a different way. They consider the bottleneck model with fixed capacity but where the FIFO property of the bottleneck model is replaced by random queue sorting, where all travellers in the queue at any given time have the same probability of exiting the queue at that moment. A range of intermediate regimes is also considered. Equations (1.10) and (1.11) still apply and the results that follow from these hence also apply.

Queues take time to dissipate. This physical property of queues has implications for how queues evolve over the course of a day. An empirical regularity of congested demand peaks is that the mean travel time peaks later than the variance of travel time. Fosgerau (2010) shows how this phenomenon arises in a dynamic model of congestion with the ratio of demand to capacity being random.

Heterogeneity

An extension to the basic bottleneck model which is clearly very important is to allow for heterogeneity. The basic model describes travellers as having identical scheduling preferences and identical preferred arrival time. This is very far from reality. For example, using survey data, Fosgerau (2006) estimates the distribution of the value of travel time, α. After conditioning on a number of controls, he finds that the remaining variation in the value of travel time has more than a factor of 50 between the 20th and 80th percentiles of the value of travel time distribution. There is every reason to think that preferences regarding earliness and lateness are similarly heterogeneous.

One of the first questions to ask when such heterogeneity is allowed in the bottleneck models is whether equilibrium still exists and whether it is unique. Analysis of the model would be severely complicated if this failed. This is the subject of Lindsey (2004), who presents general conditions under which equilibrium exists in the basic bottleneck model extended with heterogeneity in the form of a finite number of homogenous groups of travellers. Lindsey provides a review of previous literature regarding preference heterogeneity in the dynamic model.

Parking

Parking is costly in that it competes for urban space with other uses. Cruising for parking is a significant contributor to urban congestion. Arnott and co-authors have published a series of papers on this and related issues, a recent reference is Arnott and Rowse (2009).

There are a few papers on downtown parking in a dynamic framework in which parking occupies space and the attractiveness of a parking space decreases with the distance to the CBD. ADL (1991) use the bottleneck model to assess the relative efficiency of road tolls and parking fees. Without pricing, drivers occupy parking in order of increasing distance from the CBD. A time-varying toll can prevent queueing, but does not affect the order in which parking spots are taken. Optimal location-dependent parking fees may be superior; they do not eliminate queueing, but induce drivers to park in order of decreasing distance from the CBD, thereby concentrating arrival times closer to work start times. Zhang, Huang and Zhang (2008) integrate AM and PM commutes with parking in this framework.

Small networks with dynamic congestion

This section considers some simple extensions from one link to small networks. Consider first two routes in parallel connecting an origin with a destination. There are $N > 0$ travellers with $\alpha - \beta - \gamma$ scheduling preferences. They each have to choose a route and a departure time. Each route has a certain fixed travel time and a bottleneck with fixed capacity. Denote the fixed travel times by T_i and the capacities by $s_i, i = 1, 2$. Denote also the number of travellers choosing route i as $n_i > 0$ where $N = n_1 + n_2$, since all travellers choose one and only one route. Moreover, let ρ_i denote the arrival rate at the bottleneck for each of the routes.

Consider first the choice of departure time conditional on the number of travellers on each route. From the previous analysis we know that in equilibrium they incur a trip cost of $\delta n_i / s_i$ on each route. There exists a unique equilibrium where

$$\alpha T_1 + \delta \frac{n_1}{s_1} = \alpha T_2 + \delta \frac{n_2}{s_2}.$$

This is equivalent to (1.1) for the static model. It is straightforward to verify that the equilibrium number of travellers on route 1 is

$$n_1 = N \frac{s_1}{s_1 + s_2} + \frac{\alpha}{\delta} \frac{s_1 s_2}{s_1 + s_2} \left(T_2 - T_1 \right),$$

and the equilibrium cost is

$$C = \alpha \frac{s_1 T_1 + s_2 T_2}{s_1 + s_2} + \delta \frac{N}{s_1 + s_2}.$$

This shows that two bottlenecks in parallel act just like a single bottleneck. The equivalent single bottleneck would have a fixed travel time that is a weighted average of the fixed travel times on the two routes and it would have a bottleneck capacity that is the sum of the capacities of the two routes. This result can be generalised to any number of parallel routes.

A toll may be set at each bottleneck just as if it was a single bottleneck with elastic demand. As we have seen (page 6), the optimal toll does not affect the cost of using each route. Hence the split of travellers between routes is not affected by optimal tolling: The optimal toll does not reallocate between routes, but only across departure times. This is a very different conclusion than was reached in the static model, where the social optimum had a different allocation of travellers on routes than the equilibrium.

There is another situation in which several bottlenecks acts like a single bottleneck. This happens when bottlenecks are connected in a serial manner. In this case, the effective capacity is just the minimum of the bottleneck capacities. That is, the binding capacity constraint is that of the smallest bottleneck.

The property that parallel or serial bottlenecks can be reduced to a single equivalent bottleneck seems likely to survive if $\alpha - \beta - \gamma$ preferences are replaced by general preferences. The description of the equivalent bottleneck does become more complicated. The property that equilibrium usage of the parallel routes is optimal also survives.

ADL (1993b) analyse a Y-shaped network of bottlenecks to show that a Braess type paradox can arise: an increase in capacity can lead to increased cost. Analysis of more complicated networks is complicated and no general results on networks of bottlenecks seem to be available.

Large networks

The extension of the dynamic model to large networks remains a difficult problem. So far, existence and uniqueness of equilibrium have not been established (in spite of many attempts). The dynamic traffic assignment problem (Merchant & Nemhauser, 1978) is the subject of a large literature spanning several disciplines. Heydecker & Addison (2005) and Zhang & Zhang (2010) derive some analytical results.

Otherwise, the literature mostly uses numerical methods. Dynamic traffic assignment models are also difficult to work with numerically due to the dimensionality of the problem, which quickly becomes extreme.

Consider a simulation model in which travellers choose the least-cost path through a network. Conditional on the actions of all other travellers, the problem of finding the least-cost path is feasible to solve using well-established algorithms (Dijkstra, 1959). These algorithms are quite efficient but nevertheless require nontrivial time to execute. The dynamic version of such a model is formulated in continuous time; we may want to approximate it using discrete time steps of one second. In, say, a four hour peak period there are 14,400 possible departure time choices. In order to simulate the choice of departure time, we have to find the least-cost path for each possible departure time. Consider a city which can be adequately represented by a zone system of 500 zones. Then the OD-matrix, indicating the size of origin-destination flows is a 500 by 500 matrix with 25,000 entries. So the model will have to solve 3.6 billion shortest path problems through the network connecting the 500 zones. This will have to be done many times in order for such a simulation to identify an equilibrium in which no traveller will want to change his choice of departure time and route. The result is a huge computational problem and it is practically impossible to handle using a naive approach.

This section describes one approach taken to this problem, used in the model METROPOLIS (de Palma et al., 1997). The basic idea for reducing the amount of computation is to drop the assumption that travellers can choose the shortest path considering the whole network at once. At each intersection, travellers are able to observe the travel cost on each downstream link. But they do not observe the travel cost on links further downstream. Instead they are able to form an expectation regarding the travel cost from the next downstream nodes until the destination. Travellers then choose the next link with the smallest expected total cost to reach the destination, i.e. the smallest sum of the cost of the next link and the downstream expected cost. This portrays travellers as making dynamic discrete choices and these are readily formulated as a dynamic programming model using the Bellman principle.[x,xi]

The simulation model looks for equilibrium using a process which can be interpreted as a day-to-day learning process. At the end of each day, the past outcomes for all travellers are pooled and this pool of information is common knowledge. During the next day, travellers have this information available when forming expectations. The idiosyncratic error terms are the same day after day (for departure time choice model). The choice of route can be either deterministic or stochastic (in such case, error terms are i.i.d. over space and time).

Other congestion functions

Henderson (1974) formulated a dynamic model of congestion using a similar setup to Vickrey (1969), but in which the travel time is determined by the flow at the time of departure and where flows departing at different times do not interact. Chu (1995) showed that the original Henderson formulation had problems due to nonexistence of equilibrium and proposed a reformulation in which travel time for a traveller is instead determined by the flow at the time of arrival at the destination. The Chu formulation has the Vickrey bottleneck as a limiting case.

Conclusions

This paper has presented an overview of dynamic models of congestion, focusing on results derived from the Vickrey bottleneck model. This model combines in a compact way the essential

features of congestion dynamics. We have also argued that some fundamental features of congestion are inherently dynamic, which makes dynamic models indispensable for many purposes. In particular, dynamic models can be used to study a variety of policies that cannot be studied with static models. These include road pricing with a time-varying component, flexible work hours, staggered work hours, dynamic access control, and ramp metering used to differentiate capacity allocation. Pricing policies are much more effective when tolls depend on the time of the day, for stylised as well as for real networks (see Santos, 2004).

Research into congestion dynamics remains a very active area with many unresolved issues of high importance. We will mention a few here. Economic analyses using dynamic models of congestion are usually undertaken on the assumption that users are in Nash equilibrium. It would therefore be of interest to give general conditions under which Nash equilibrium exists (for general networks). It would further be of interest to specify learning mechanisms that would lead to Nash equilibrium. A learning mechanism is a rule that travellers use to update their choice of departure time and route in the presence of information concerning past outcomes. The existence of learning mechanisms leading to Nash equilibrium would support the presumption that the notion of Nash equilibrium is useful as a benchmark for actual congestion phenomena. Knowledge about learning mechanisms leading to Nash equilibrium may also be useful for the design of algorithms to find Nash equilibrium in simulation models.

Progress would also be desirable concerning the nature of scheduling preferences. The discussion in this paper has taken for granted that travellers are equipped with scheduling preferences and that these can be regarded as exogenous from the point of view of our analysis. Our transportation perspective has led us to be concerned with the timing of trips and we view travellers simply as having preferences regarding timing such that they can respond to circumstances by changing their trip timing in sensible ways. These times are hardly the fundamental objects of preference and, strictly speaking, it only makes sense to formulate preferences in these terms when circumstances such as the activities before and after the trip can be regarded as exogenous. This is, however, not a very appealing position. If I know that my trip will take more time, then I will adjust my schedule for the day to take this into account. I care, e.g., about not being late for appointments. But I make the appointments myself and so my scheduling preferences are a consequence of choice.

It is natural to ask why commuters mostly prefer to arrive at work at the same time. Various contributions have answered this question by pointing to agglomeration forces at the workplace, whereby productivity and wages are affected by the degree of overlap in work times, see Henderson (1981), Wilson (1988), Hall (1989). If this view is correct, then changes to the transport system will affect agglomeration, which in turn will affect commuter scheduling preferences. It remains to be seen how such mechanisms matter for our understanding of the effect of transport policies.

Endogeneity of scheduling preferences may also matter for the value of information. Consider a trip exposed to random travel time variability. At some point in time I will learn the size of delay. If scheduling preferences are exogenous, then it only matters whether I learn about the size of the delay soon enough to adjust my departure time. If scheduling preferences are endogenous, then it also matters whether I learn about the size of the delay soon enough to adjust my schedule (Kreps and Porteus (1978) consider dynamic choice behaviour under conditions of uncertainty, with emphasis on the timing of the resolution of uncertainty.

As discussed in this paper, the current state of the topic of dynamic congestion modelling provides a range of general insights from small stylised models. Numerical simulation models exist to deal with the complexities of real size networks. In between, there is a large gap. Numerical simulation has the drawback that it must rely on particular assumptions, which may or may not provide good approximations to the object of interest. So a main motivation for continued theoretical research into dynamic models of congestion is the desire for increased generality. The fewer assumptions required for a conclusion, the more certain we can be that it applies. As this paper has discussed, there are a number of directions in which we would like to extend our models so that they become better able to account for the facts that travellers are very heterogeneous, they make route and scheduling decisions based on limited information, they interact heavily in ways related to scheduling and they move about in complex networks that are subject to random shocks. The other main motivation for research into the area is the potential for providing a better empirical foundation for our models. One possibility that naturally comes to mind is to seek to utilise data sources such as GPS data to obtain a better understanding of actual trip scheduling behaviour.

In conclusion, many exciting things have been done, giving us many important insights into congestion dynamics, and there are still many exciting things waiting to be done.

Reference List

Aguirregabiria, V. & Mira, P. (2010). Dynamic discrete choice structural models: A survey. *Journal of Econometrics, 156,* 38-67.

Arnott, R. A., de Palma, A., & Lindsey, R. (1990). Economics of a bottleneck. *Journal of Urban Economics, 27,* 111-130.

Arnott, R.A., de Palma, A., & Lindsey, R. (1991). A temporal and spatial equilibrium analysis of commuter parking. *Journal of Public Economics, 45,* 301-335.

Arnott, R. A., de Palma, A., & Lindsey, R. (1993a). A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *American Economic Review, 83,* 161-179.

Arnott, R.A., de Palma, A., & Lindsey, R. (1993b). Properties of Dynamic Traffic Equilibrium Involving Bottlenecks, Including a Paradox and Metering. *Transportation Science, 27,* 148-160.

Arnott, R. A., de Palma, A., & Lindsey, R. (1999). Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand. *European Economic Review, 43,* 525-548.

Arnott, R. & Rowse, J. (2009). Downtown parking in auto city. *Regional Science and Urban Economics, 39,* 1-14.

Beckmann, M. J., McGuire, C. B., & Winston, C. B. (1956). *Studies in the Economics of Transportation*. New Haven, Connecticut: Yale University Press.

Chu, X. (1995). Alternative congestion pricing schedules. *Regional Science and Urban Economics, 29,* 697-722.

Daganzo, C. F. & Garcia, R. C. (2000). A Pareto Improving Strategy for the Time-Dependent Morning Commute Problem. *Transportation Science, 34,* 303-311.

de Palma, A., Ben-Akiva, M., Lefevre, C., & Litinas, N. (1983). Stochastic Equilibrium Model of Peak Period Traffic Congestion. *Transportation Science, 17,* 430-453.
de Palma, A. & Fosgerau, M. (2009). Random queues and risk averse users. *Working Paper*, Ecole Polytechnique, France.
de Palma, A., Lindsey, R., & Proost, S. (2007). Investment and the use of tax and toll revenues in the transport sector: The research agenda. In *Investment and the Use of Tax and Toll Revenues in the Transport Sector* (pp. 1-26).
de Palma, A., Marchal, F., & Nesterov, Y. (1997). METROPOLIS - Modular System for Dynamic Traffic Simulation. *Transportation Research Record, 1607,* 178-184.
Dijkstra, E. W. (1959). A note on two problems in connection with graphs. *Numerische Mathematik, 1,* 269-271.
Fosgerau, M. (2006). Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological, 40,* 688-707.
Fosgerau, M. (2009). The marginal social cost of headway for a scheduled service. *Transportation Research Part B: Methodological, 43,* 813-820.
Fosgerau, M. (2010). On the relation between the mean and variance of delay in dynamic queues with random capacity and demand. *Journal of Economic Dynamics and Control, 34,* 598-603.
Fosgerau, M. & Engelson, L. (2010). The value of travel time variance. *Transportation Research Part B, Forthcoming.*
Fosgerau, M. & Karlstrom, A. (2010). The value of reliability. *Transportation Research Part B, 44,* 38-49.
Hall, R. E. (1989). Temporal agglomeration. *National Bureau of Economic Research, 3143.*
Haurie, A. & Marcotte, P. (1985). On the relationship between Nash-Cournot and Wardrop equilibria. *Networks, 15,* 295-308.
Henderson, J. V. (1974). Road congestion : A reconsideration of pricing theory. *Journal of Urban Economics, 1,* 346-365.
Henderson, J. V. (1981). The economics of staggered work hours. *Journal of Urban Economics, 9,* 349-364.
Heydecker, B. G. & Addison, J. D. (2005). Analysis of Dynamic Traffic Equilibrium with Departure Time Choice. *Transportation Science, 39,* 39-57.
Kreps, D. M. & Porteus, E. L. (1978). Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica, 46,* 185-200.
Laih, C. H. (2004). Effects of the optimal step toll scheme on equilibrium commuter behaviour. *Applied Economics, 36,* 59-81.
Laih, C.-H. (1994). Queueing at a bottleneck with single- and multi-step tolls. *Transportation Research Part A, 28,* 197-208.
Lindsey, R. (2004). Existence, Uniqueness, and Trip Cost Function Properties of User Equilibrium in the Bottleneck Model with Multiple User Classes. *Transportation Science, 38,* 293-314.
Lindsey, R. (2009). Cost recovery from congestion tolls with random capacity and demand. *Journal of Urban Economics, 66,* 16-24.
Merchant, D. K. & Nemhauser, G. L. (1978). A Model and an Algorithm for the Dynamic Traffic Assignment Problems. *Transportation Science, 12,* 183-199.
Mohring, H. & Harwitz, M. (1962). *Highway Benefits: An Analytical Framework.* Evanston, Illinois: Northwestern University Press.
Santos, G. (2004). Research in Transportation Economics, 9, XI-XIII.

Shen, W. & Zhang, H. M. (2010). Pareto-improving ramp metering strategies for reducing congestion in the morning commute. *Transportation Research Part A: Policy and Practice, 44,* 676-696.
Small, K. (1982). The scheduling of Consumer Activities: Work Trips. *American Economic Review, 72,* 467-479.
Tseng, Y. Y. & Verhoef, E. T. (2008). Value of time by time of day: A stated-preference study. *Transportation Research Part B: Methodological, 42,* 607-618.
Verhoef, E. T. & Mohring, H. (2009). Self-Financing Roads. *International Journal of Sustainable Transportation, 3,* 293-311.
Vickrey, W. S. (1969). Congestion theory and transport investment. *American Economic Review, 59,* 251-261.
Wardrop, J. G. (1952). Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institute of Civil Engineering, Part II,* 325-378.
Wilson, P. W. (1988). Wage variation resulting from staggered work hours. *Journal of Urban Economics, 24,* 9-26.
Zhang, X. & Zhang, H. (2010). Simultaneous Departure Time/Route Choices in Queuing Networks and a Novel Paradox. *Networks and Spatial Economics, 10,* 93-112.

---

[i] We would like to thank Robin Lindsey for many useful comments, suggestions and references. The first author would like to thank the Institut Universitaire de France.

[ii] For a more detailed analysis of congestion in the static model see the G. Santos and E. Verhoef contributions.

[iii] We assume the parameters are such that this equation leads to positive flows on each route.

[iv] C' denotes the derivative of C.

[v] Small (1982) tested a range of formulations of scheduling preferences, including the $\alpha-\beta-\gamma$ preferences as a special case.

[vi] Since $c_2(t_1,t_2)<0$. This follows since $\beta<\alpha$. We use subscripts to denote partial derivatives.

[vii] We use subscripts to denote partial derivatives.

[viii] Provided that the toll does not decrease too quickly. A quickly decreasing toll may induce travelers to avoid certain departure times, which leads to unused capacity.

[ix] Laih (1994) did not recognize that it was necessary to reformulate the queueing technology in order to obtain his results. This was rectified in Laih (2004).

[x] The exact optimization procedure used in METROPOLIS was never published since it is commercial proprietary software.

[xi] Dynamic discrete choice models are surveyed in Aguirregabiria & Mira (2010).